Semantic Analysis for Paraphrase Identification using Semantic Role Labeling

Eunji Lee
Chosun University
eunbesu@gmail.com

Htet Myet Lynn
Chosun University
htetmyet@chosun.ac.kr

Hyoungju Kim
Chosun University
snowlisakim@gmail.com

Soonja Yeom
University of Tasimania
soonja.yeom@utas.edu.au

Pankoo Kim
Chosun University
pkkim@chosun.ac.kr

ABSTRACT

Reuse of documents has been prominently appeared during the course of digitalization of information contents owing to the wide-spread of internet and smartphones in various complex forms such as inserting words, omitting and substituting, changing word order, and etc. Especially, when a word in document is substituted with a similar word, it would be an issue not to consider it as a subject of measurement for the existing morphological similarity measurement method. In order to resolve this kind of problem, various researches have been conducted on the similarity measurement considering semantic information. This study is to propose a measurement method on semantic similarity being characterized as semantic role information in sentences acquired by semantic role labeling. To assess the performance of this proposed method, it was compared with the method of substring similarity being utilized for similarity measurement for existing documents. As a result, we could identify that the proposed method performed similar with the conventional method for the plagiarized documents which were rarely modified whereas it had improved results for paraphrasing sentences which were changed in structure.

CCS CONCEPTS

• **Information Systems** → Similarity measures; • **Applied Computing** → Document analysis;

KEYWORDS

Text Reuse, Text Similarity, Semantic Role Labeling, PAN 2012 Corpus

ACM Reference format:

Eunji Lee, Htet Myet Lynn, HJ Kim, and Pankoo Kim, Semantic Analysis for Paraphrase Identification using Semantic Role Labeling. In *Proceedings of ACM SAC Conference, Limassol, Cyprus, April 8-12, 2019,* (SAC'19), 4 pages. DOI: https://doi.org/10.1145/3297280.3297623

SAC '19, April 8–12, 2019, Limassol, Cyprus

ACM ISBN 978-1-4503-5933-7/19/04.
https://doi.org/10.1145/3297280.3297623

1 INTRODUCTION

As the volume of information is exponentially increased along with the fact that the various information including newspaper articles, books, and academic papers has been digitalized and existed online leading to easy and convenient access to the information, the social problems such as unauthorized use and plagiarism has been emerged at the same time due to indiscreet sharing [1]. Plagiarism in this era of information has adversely affect desirable distribution and utilization, thus the researches on measurement of document similarity to more efficiently detect plagiarism. The measurement methods of document similarity can be classified into 2 categories, for instance, comparing morphological similarity and comparing semantic similarity. For comparing morphological similarity, some of the representative methods include n-gram method, in which extracting and comparing adjacent 'n' number of words in a sentence, comparing substring in a sentence, and VSM(Vector Space Model), in which similarity is determined by measuring the distance between vectors after putting a document in a vector space [2]. However in this morphological similarity comparison, it has a drawback that similarity measurement does not consider for the case when the original document has been altered such as substituting, paraphrasing, and rewriting words since it only compares words in two subject documents [2]. The semantic similarity measurement method was proposed to improve this kind of issues. The semantic similarity measurement method is a method using knowledge base and thesaurus, in which semantic relationships among words are defined in hierarchy information. It may be useful for only measuring similarity of words in a sentence considering semantic similarity with substituting or changing to similar words, but it cannot comprehend the structural information in a sentence where the subject words are at. Therefore, it does have limitation on detecting types of plagiarism which evolves in various forms including paraphrasing sentences [3, 4]. Most people think that copying the expression from other writings is plagiarism but paraphrasing, which borrows and alters the expression, is not. Despite the fact that expression is borrowed and changed, it would be plagiarism if the entire message and structure is the same. This study is to propose the methods of measuring similarity between

documents utilizing SRL(Semantic Role Labeling), which is one of the semantic analysis methods based on structural information of sentence in order to measure similarity of documents similar in semantic being composed of much altered from the original such as paraphrasing sentences.

2 RELATED STUDIES

2.1 Semantic Role Labeling

The semantic role labeling analyzes the sentence components as predicate-argument structure based on sentence structure analysis, and determines and tags the semantic role in a sentence for each component of sentence [5, 6]. The semantic role labeling identifies argument information required for completing a sentence by predicates centering of natural language sentences, and determines the relationships between predicates and corresponding arguments. By mapping the semantic role arguments with acting agent, experiencing agent, object, and etc., semantic role labeling can be conducted [7]. The predicate-argument structure of sentence is an important component to represent the semantic of sentence while predicate of particular semantic needs essential argument information. This will result in using common predicate-argument information for sentences of similar semantic. The conduct of semantic role labeling can be divided into the following 4 stages: PI(Predicate Identification) stage, PC(Predicate Classification) stage, AI(Argument Identification) stage, and AC(Argument Classification) stage [8]. Firstly the PI stage, where predicate is identified through sentence structure analysis on input sentence, and the PC stage, where predicate is ambiguity of predicate can be resolved, are conducted, then the AI stage, where information required for making a sentence with predicates with particular semantic in the sentence is identified, and the Ac stage, where semantic role of acknowledged argument is determined, are conducted [8]. In this study, the semantic role labeling is utilized to improve detection performance on similar sentences being much altered by paraphrasing sentences, which are difficult to detect with conventional similarity measurement methods.

2.2 Knowledge Bases for Semantic Role Labeling

The representative Knowledge Bases for Semantic Role Labeling include Propbank and FrameNet. They provide the corpuses tagged with semantic role [9]. The Propbank (Martha Palmer, 2005) is constructed in a way that semantic role are added by argument in syntactic structure analysis database of Penn Tree bank. The Propbank considers only verbs as predicates in order to establish the relationship between predicate and argument [10]. On the other hand, the FrameNet establishes predicate-argument information for corpuses such as BNC(British national Corpus) and ANC(American National Corpus) while it is established as a frame-argument information structure considering not only verbs but also argument relationship including noun, adverb, and adjective by applying 'Frame' the classification concept [11, 12].

3 SEMANTIC SIMIARITY MEASUREMENT USING SEMANTIC ROLE LABELING

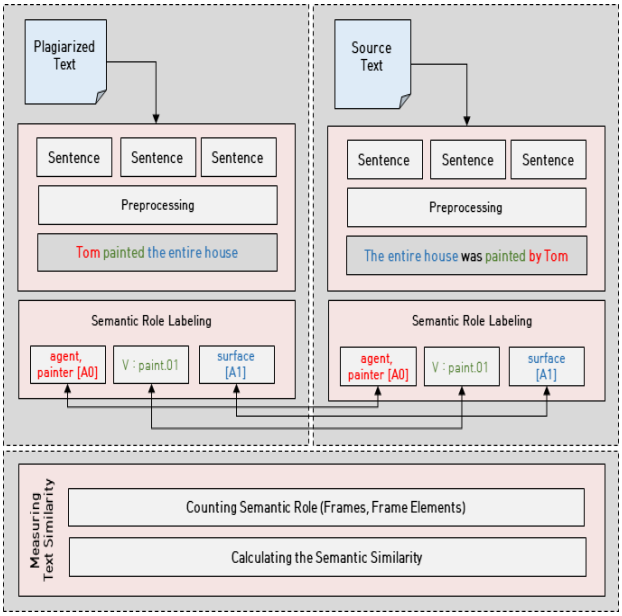

Figure 2: Structure for Proposed Method

In this session, the semantic role labeling of document is conducted by using FrameNet, the measurement method of semantic similarity of document by comparing the semantic role information in the sentence is described. The document similarity measurement process using semantic role labeling is composed of pre-process, semantic role labeling process, and semantic similarity measurement process. Fig. 2 is a structure for document similarity measurement method by using FrameNet.

3.1 Preprocessing

The sentence segmentation, stopword removal, and lemmatization are conducted in the pre-process. The sentence segmentation is a process to segment a document into single sentences to conduct morpheme analysis or sentence structure analysis as a pre-process. The proposed method in this study is to conduct the sentence segmentation process segmenting the input document into sentence units since it conducts semantic role labeling based on the sentence structure analysis for similarity measurement of documents. In this study for stopword removal, it is intended to increase the processing speed of the system by utilizing pre-defined stopwords and removing all stopwords in the text. The lemmatization is a process to find lemma for altered words of sentences in various forms. The lemmatization is appraised as an efficient method to resolve the problem with the fact that the same semantic words in an altered form are identified as different information in the document analysis [13]. In this study, we intend to improve the performance of sentence structure analysis by lemmatization.

3.2 Similarity Measurement Using Semantic Role Labeling

The document similarity measurement is to measure similarity of documents by comparing the document as a test object and the base document while it is generally conducted in a way of selecting a candidate sentence from sets of documents to compare by using guide words in sentences, and measuring the similarity of documents by similarity measurement between the two sentences. The proposed method is to tag words in sentences with the relevant semantic role and to focus on the similarity measurement between the semantic role of two sentences while arguments of similar semantic in sentences with similar contents can appear in semantic role since argument with semantic relationship from predicate are distributed in sentences. Comparing similarity based on the semantic role in the sentence can reduce time and cost to be consumed for similarity measurement of sentences, and it has a merit to detect paraphrasing type which includes the same semantic though the sentence is altered. The process for tagging semantic role in this study used SEMAFOR, open source API developed for semantic analysis of documents. The result of semantic role labeling conducted by the SEMAFOR system creates JSON file being tagged with frame-argument information. The table 1 and table 2 intuitively illustrates the semantic role information of sentences. Since the two sentences of the table 1 and table 2 have the same semantic yet paraphrased in different form, we can identify that they have the common semantic role information. The similarity of the two sentences is calculated by Greedy String Tiling similarity, [14] which is a comparison method utilizing substring.

Table 1: Semantic Role Labeling Information of Sentence #1 (Source_Document_01000.txt (PAN2012 corpus))

Sentence #1	**The cry of, Pig out! and the consequent rush of children in pursuit, at last reached such a pitch that both Miss Grey and the much-tried Andrew made complaint to the vicar.**	
Index	***frames_name***	***frameElements_name***
1	Vocalizations	-
2	Locative_relation	-
3	**Self_motion**	-
4	Kinship	Alter
5	**Seeking_to_achieve**	-
6	**Relative_time**	**Focal_occasion**
7	Arriving	Goal
8	-	Theme
9	**Quantity**	**Quantity**
10	Success_or_failure	-
11	**Causation**	**Cause**
12	-	**Effect**
13	Complaining	-

Table 2: Semantic Role Labeling Information of Sentence #1 (Suspicious_Document_01000.txt (PAN2012 corpus))

Sentence #1	**The call of, "Pig away!" and the dash of bairn in the pursuit, at last make such a soprano that both attend grey and the much-try Andrew make disorder to the vicar.**	
Index	***frames_name***	***frameElements_name***
1	Request	-
2	**Self_motion**	-
3	**Seeking_to_achieve**	-
4	**Relative_time**	**Focal_occasion**
5	Type	Subtype
6	-	Category
7	-	Type_Property
8	**Quantity**	**Quantity**
9	Attending	Agent
10	-	Event
11	**Causation**	**Cause**
12	-	**Effect**
13	Medical_conditions	-

In this method, the number of semantic role labeling is defined as characteristic information instead of the number of substring while the similarity of the two sentences is calculated based on the ratio of characteristic information found commonly between the two sentences. The similarity is calculated as the formula 1.

$$\text{SRL Similarity} = \frac{2*\sum_{i \in srl} length_i}{|Srl_{sus}+Srl_{src}|} \quad (1)$$

4 EXPERIMENT

4.1 Similarity Measurement

In this session, the semantic role labeling on experimental data for similarity measurement is conducted by using the FrameNet. For the experiment, the proposed method and the conventional similarity measurement method are applied on 500 documents being composed of source document and suspicious document in pairs within the PAN 2012 corpus (simulated_paraphrase category). The table 3 is the result of measurement of similarity by substring comparison method on the experimental data set whereas the table 4 is the result of the result of measurement of similarity on the experimental data set by the semantic role labeling information.

Table 3: The Results of Substring Similarity

pair	***# source text***	***# suspicious text***	***# match word***	***Substring Similarity***
00000	888	784	172	0.2057416
00001	93	94	70	0.7486631
00002	164	135	42	0.2809365
...	...	...	...	...

00497	324	331	199	0.6076336
00498	42	32	7	0.1891892
00499	213	195	39	0.1911765

Table 4: The Results of Semantic Role Labeling Similarity

pair	*# source text*	*# suspicious text*	*# match semantic roles*	*SRL Similarity*
00000	667	613	281	0.439063
00001	57	58	37	0.643478
00002	86	72	27	0.341772
...	...	...	...	...
00497	232	236	140	0.598291
00498	41	37	25	0.641026
00499	131	140	76	0.560886

Table 5: Comparison on Similarity Measurement

Category	*Substring similarity*	*Semantic Role Labeling similarity*
simulated_paraphrase	0.21067	0.53742

4.2 Evaluation

The proposed accuracy rate and recall factor are calculated by the formula 2 and 3. The 'S' is the range of plagiarism which is 'correct group' to assess the proposed method while 'R' is the range of plagiarism found by the proposed method. The 'r' is the number of common semantic role information from the two documents extracted by the proposed method whereas the 's' is the number of semantic role within the plagiarism range, 'r∩s' is calculated by the number of semantic role within the plagiarism range among the number of common semantic role information from the two documents. The table 6 illustrates accuracy rate and recall factor measured on the similarity between documents based on the tagged semantic role labeling information through the extended FrameNet.

$$\mathrm{pre}(S, R) = \frac{1}{|R|}\sum_{r\in R}\frac{\cup_{s\in S}(s\cap r)}{|r|} \quad (2)$$

$$\mathrm{rec}(S, R) = \frac{1}{|S|}\sum_{s\in S}\frac{\cup_{s\in S}(s\cap r)}{|s|} \quad (3)$$

Table 6: Performance of Proposed Method

Plagiarism correct # of semantic role (A)	*# semantic role extracted (B)*	*Plagiarism correct coincident with # of semantic role (C)*	*Precision (C/B)*	*Recall (C/A)*
18303	9278	5724	0.61692	0.31273

5 CONCLUSIONS

This study proposed the semantic similarity measurement method through structure analysis of sentences using semantic role labeling whereas the semantic role labeling is the method of document analysis to determine the role of argument information which is required for completing a sentence being composed of predicate centering on the predicate of sentence. The semantic role labeling information acquired through this process as a characteristic of document is used for the similarity measurement. It is a method with especially not only semantic of words but also structure of document is considered leading to identify more improved performance than the conventional similarity measurement method on paraphrased documents. As a future study, we plan to research further on similarity measurement methods being applied by expansion measures being able to enrich semantic role information of existing FrameNet.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIP) (No. NRF-2016R1A2B4012638).

REFERENCES

[1] Leuf, B. and Cunningham, W. 2001. The Wiki Way: Collaboration and Sharing on the Internet.

[2] Daniel, B., Torsten, Z. and Iryna, G. 2012. Text Reuse detection using a composition of text similarity measures. In *Proceeding of the 23rd International Conference on Computational Linguistics* (COLING '12). Bombay, India,167-184.

[3] Mihalcea, R., Corley, C. and Strapparava, C. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence* (AAAI' 06). Boston, USA, 775–780.

[4] Sumathy. K. L. and Chidambaram. 2016. A Hybrid Approach for Measuring Semantic Similarity between Documents. *International Journal of Advanced Computer Science and Applications*, Vol. 7, No. 8. 231-237. DOI: 10.14569/IJACSA.2016.070831

[5] Liu, D. and Gildea, D. 2010. Semantic Role Features for Machine Translation. In *Proceeding of the 21st International Conference on Computational Linguistics* (COLING '10). Sydney, Australia, 716-724.

[6] Woodsend, K. and Lapata, M. 2017. Text Rewriting Improves Semantic Role Labeling. In *Proceedings of the 31st National Conference on Artificial Intelligence* (AAAI' 17). San Francisco, USA, 5095-5099.

[7] Yakushiji, A., Miyao, Y., Ohta, T., Tateisi, Y. and Tsujii, J. 2006. Automatic Construction of Predicate-argument Structure Patterns for Biomedical Information Extraction. In *Proceeding of the Conference on Empirical Methods in Natural Language Processing* (EMNLP' 06). Sydney, Australia, 284-292.

[8] Das, D., Chen, D., Martins, A. F. T., Schneider, N. and Smith, N. A. 2014. Frame-Semantic Parsing. *Journal of Computational Liguistics*, Vol. 40, Issue. 1. 9-56. DOI: https://doi.org/10.1162/COLI_a_00163

[9] Shi, L. and Mihalcea, R. 2005. Putting Pieces Together : Combining FrameNet, VerbNet, and WordNet for Robust Semantic Parsing. Lecture Notes in Computer Science, Vol. 3406. Springer-Verlag, London, 100–111.

[10] Palmer, M. and Gildea, D. 2005. The Proposition Bank : An Annotated Corpus of Semantic Roles. *Journal of Computational Linguistics*, Vol.31, Issue. 1. 71-106. DOI: 10.1162/0891201053630264

[11] Fillmore, C. J. and Baker, C. 2009. A frames approach to semantic analysis. In Bernd Heine and Heiko Narrog, editors, The Oxford Handbook of Linguistic Analysis. 791-816.

[12] https://framenet.icsi.berkeley.edu/fndrupal/frameIndex

[13] Manning, C. D., Raghavan, P. and Schütze, H. 2008. Introduction to Information Retrieval. Cambridge University Press.

[14] Kumar, J. A. 2012. Similarity Overlap Metric and Greedy String Tiling at PAN 2012: Plagiarism Detection. In *2012 Conference and Labs of the Evaluation Forum* (CLEF' 12). 1-6.